Short-Term Ensemble Streamflow Prediction Using Spatially Shifted QPF Informed by Displacement Errors

KYLE K. HUGEBACK,[a] KRISTIE J. FRANZ,[a] AND WILLIAM A. GALLUS JR.[a]

[a] *Department of Geological and Atmospheric Science, Iowa State University, Ames, Iowa*

(Manuscript received 5 January 2022, in final form 25 October 2022)

ABSTRACT: Errors associated with the location of precipitation in QPFs present challenges when used for hydrologic prediction, particularly in small watersheds. This work builds on a past study that systematically shifted QPFs prior to inputting them into a hydrologic model to generate streamflow ensembles. In the original study, which used static, predetermined shifting distances, flood detection improved, but false alarms increased due to large ensemble spread. The present research tests a more informed approach by randomly selecting shift directions and distances based on the distribution of displacement errors from a sample of QPFs. Precipitation forecasts were taken from the High-Resolution Rapid Refresh Ensemble (HRRRE), and streamflow predictions were generated using the Weather Research and Forecasting hydrological modeling system, version 5.1.1, in a National Water Model 2.0 configuration. A 63-member streamflow ensemble was generated using the 9 original HRRRE and 54 shifted HRRRE members. Two ensemble updating schemes were tested in which ensemble member weights were adjusted using precipitation location and QPF displacement present at convective initiation. The ensembles using QPF shifted based on climatological spatial errors showed higher probabilistic forecasting skill, while having comparable dichotomous forecasting skill to the original HRRRE ensemble. Other methods of selecting nine ensemble members from the full 63-member suite did not show significant improvement. Flood peak timing showed frequent errors, with average timing errors around five hours early. Larger watersheds tended to have better skill metric scores than smaller basins, with increased skill added by the shifting of QPF.

KEYWORDS: Ensembles; Numerical weather prediction/forecasting; Hydrologic models; Flood events

1. Introduction

Since 1980, there have been 35 separate billion-dollar flood disasters in the United States, with a total cost exceeding $159 billion. Additionally, flooding during that time accounted for 624 fatalities (Smith 2020). In much of the central United States, a majority of the annual precipitation falls during the warm months of the year from May to September (Shaw and Waite 1964). The warm season convective rainfall in this region is dominated by large mesoscale convective systems (MCSs) that pose a number of different hazards, including severe wind, hail, tornadoes, and intense localized precipitation (Fritsch et al. 1986; Gallus 2012; Haberlie and Ashley 2019). MCSs account for over 50% of annual rainfall in the Midwest and eastern United States (Haberlie and Ashley 2019). Future climate projections have indicated a risk of more intense and frequent convective precipitation events (Hejazi and Markus 2009; Andresen et al. 2012). Given their regularity and intensity, the ability to accurately predict these events and model their hydrologic impacts is extremely important.

Errors in QPFs have been found to be the highest during the warm season when intense rainfall and flooding is most common (Gallus 2012; Sukovich et al. 2014). Errors in the predicted rainfall intensity, orientation, size, shape, timing, and location can propagate and interact with errors in the hydrologic model (Brown and Heuvelink 2006; Collier 2007), impacting the accuracy of streamflow predictions, especially for flash flooding (Rezacova et al. 2007; Hapuarachchi et al. 2011). Furthermore, uncertainty in the streamflow forecasts tends to grow with longer lead times (Seo et al. 2018; Lin et al. 2005). Past research has found that streamflow forecasts for small watersheds had significant increases in error beyond a forecast window of 6 h (Seo et al. 2018; Adams and Dymond 2019).

Gallus (2010) examined QPF skill of a 5-member, convection-allowing ensemble using object-oriented verification approaches and found there was less than a 10% error in forecasted rain rate, but the average spatial displacement was between 100 and 250 km, depending on the approach used. Over half of the systems studied had no overlap between the observations and the simulated precipitation objects, with spatial displacement errors having large standard deviations. The standard deviations ranged between 100 and 200 km, depending on which method was used to identify rainfall objects (Gallus 2010). A recent study of a 2018 flood event showed that QPF intensities from the High-Resolution Rapid Refresh (HRRR) model were accurately depicting the risk of flooding in the mesoscale study domain, while errors in the spatial displacement of precipitation were common at finer scales (Viterbo et al. 2020). Specifically, the average precipitation over a large area was correct, but localized areas of wet biases next to dry biases showed the impact of spatial displacement errors.

Carlberg et al. (2020) tested an ensemble streamflow prediction method that accounted for uncertainty due to spatial

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-21-0252.s1.

Corresponding author: Kyle Hugeback, hugeback@iastate.edu

DOI: 10.1175/JHM-D-21-0252.1

displacement errors by shifting ensemble member QPF prior to input into a hydrologic prediction model. They used High-Resolution Rapid Refresh Ensemble (HRRRE) QPF as forcing and shifted each of the 9 original HRRRE members in both the cardinal and ordinal directions, resulting in an 81-member ensemble when including the original HRRRE members. Shift distances of 55.5 and 111 km and 0.5° and 1.0° latitude/longitude, respectively, were tested. The probability of detection for flooding was improved in the shifted ensemble compared to the original 9-member ensemble. However, despite the shifting, often only a small subset of the ensemble members produced rainfall over the basin, leading to low probabilities associated with flooding.

Kiel et al. (2022) quantified and compared the spatial displacement errors in the High-Resolution Ensemble Forecast, version 2, (HREF) and the HRRRE at the hour of convective initiation (CI) and for the 0–18-h forecast period when most of the precipitation occurred in the systems. The displacement analysis was based on the centroids of the precipitation area and conducted for 30 cases over the 2018 warm season. Displacement errors were generally larger and had higher spread at CI than for the 0–18-h accumulation period. Location errors in the HRRRE had smaller spread and a westward bias that appeared to follow a skewed-normal spatial distribution, while HREF had larger spread and no apparent displacement directional bias (Kiel et al. 2022). In addition, Kiel et al. (2022) found that there was a slight correlation between the displacement of the forecasted precipitation centroid at CI and the centroid of the 0–18-h accumulation period. They suggest that it may be possible to improve QPF forecasts using the displacement error present at the onset of rainfall. Supporting the idea that climatological information about typical displacement errors can be helpful in adjusting forecasts, Carlberg et al. (2020) found that the ranked probability score of the ensembles improved when applying a simple weighting scheme to account for the trend of westward bias in the HRRRE precipitation.

In the present study, we test an ensemble shifting method similar to the approach of Carlberg et al. (2020), but one that is informed by the findings of Kiel et al. (2022). The HRRRE QPF fields were shifted using a randomly sampled direction and distance selected from the climatology of displacements identified by Kiel et al. (2022). The shifted QPF along with the original QPF were input into the Weather Research and Forecasting hydrological modeling system (WRF-Hydro), version 5.1.1, in the National Water Model, version 2.0, (NWM) configuration (Gochis et al. 2020) to generate ensemble streamflow predictions for 50 stream gauges over 29 events from the 2018 warm season. The forecast region covers the same area of the north-central United States that was used in related forecasting studies (Carlberg et al. 2020; Goenner et al. 2020; Kiel et al. 2022). Following Kiel et al. (2022), weighting schemes were tested to adjust the streamflow ensembles based on the displacement errors present in rainfall centroids at CI, a method that could be used to adjust ensembles in real-time forecasting. Both dichotomous and probabilistic forecasts of peak streamflow from the ensembles were verified. While the present study represents a scenario in which the displacement of the QPF members is perfectly known, it demonstrates the extent to which shifting QPF in ensembles could enhance ensemble streamflow prediction.

2. Data and methods

a. Models and datasets

The QPFs used in this study were from the 9-member HRRRE (Dowell et al. 2018; Dowell 2020), the same QPFs used in Carlberg et al. (2020) and Kiel et al. (2022). The experimental HRRRE forecasts were obtained from the NOAA Global Systems Laboratory using their FTP server. At the time data were compiled for this study, the HRRRE's domain covered the eastern two-thirds of the contiguous United States (CONUS), with the model core HRRR, version 3 (HRRRv3; Dowell et al. 2018). Lateral boundary conditions (LBCs) for the HRRRE are provided by the Rapid Refresh model, and data assimilation is accomplished using the NOAA Gridpoint Statistical Interpolation analysis system (Benjamin et al. 2016). Initial condition (IC) and LBC perturbations are added to the pressure, temperature, wind vectors **U** and **V**, and water vapor mixing ratio to create ensemble members. The Smirnova/Rapid Update Cycle land surface scheme was used with the MYNN planetary boundary layer scheme and the Thompson microphysics scheme (Smirnova et al. 2016; Nakanishi and Niino 2009; Thompson and Eidhammer 2014).

Data fields from the North American Land Data Assimilation System, version 2 (NLDAS-2; Xia et al. 2012a,b) were used as input to the WRF-Hydro when it was not in forecast mode (i.e., for model spin up). NLDAS-2's domain covers the CONUS at 0.125° spacing. Observed data from the National Centers for Environmental Prediction stage II Doppler radar precipitation estimates and the Climate Prediction Center unified gauge-based precipitation data are disaggregated to the hourly time scale and assimilated onto the grid of NLDAS-2 (Xia et al. 2012a). Preprocessing tools were used to regrid NLDAS-2 data to the parent grid of the hydrologic model. NLDAS-2 data from October 2013 to April 2018 were used to create a 4-yr spinup for the hydrologic model. NLDAS-2 was also used to create warm starts for each HRRRE storm event.

The hydrologic model used was the WRF-Hydro within the NWM (Gochis et al. 2020). A subset of the NWM for the study region (Fig. 1) and with all the necessary configurations and parameterizations was obtained from the National Center for Atmospheric Research (NCAR). Catchments and river vectoring in the NWM are derived from National Hydrography Dataset Plus (NHDPlus), version 2, data (Gochis et al. 2020). The NWM uses the Noah multiparameterization (NoahMP; Niu et al. 2011) LSM, run on a 1-km horizontal grid with a 2-m soil depth split into four layers and operated on an hourly time step. Land-cover parameters used in the NoahMP were classified using the United States Geological Survey (USGS) 24-type land use–land cover product and the MODIS modified IGBP 20-category land-cover product, and soil classifications and soil hydraulic parameters are

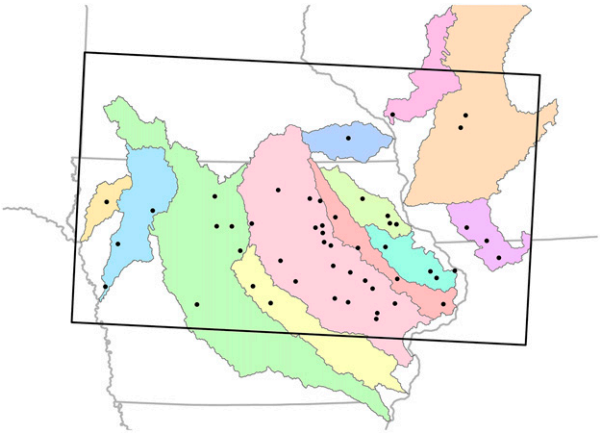

FIG. 1. NWM domain obtained from NCAR (bolded rectangle). Gauges are indicated by black dots with major watersheds shown in colored polygons.

based on the 1-km State Soil Geographic (STATSGO) Database (Gochis et al. 2020). Within the NoahMP, vegetation dynamics are calculated using climatological lookup tables. Canopy stomatal resistance is derived through the Ball–Berry conductance model (Ball et al. 1987), while the soil moisture factor for stomatal resistance comes from the Noah model. Radiative transfer in the vegetive layer is based on two-stream approximation applied to the vegetated fraction. The temperatures at the bottom of the four soil layers are taken from climatology (Niu et al. 2011; Gochis et al. 2020).

Surface runoff in the NWM setup was routed using the steepest descent option run on a 250-m nested grid at 10-s time steps. Groundwater was modeled using the built-in exponential bucket model. Water exiting the groundwater bucket is routed directly to the stream network. Last, Muskingum–Cunge reach-based routing was used to convey water downstream after it entered the channel; the reach-based routing runs at a 300-s time step, and discharge is output at hourly time steps for specified stream locations. The HRRR regridding package, provided by NCAR, was used to isolate the study domain from the much larger HRRRE domain and scale the 3-km HRRRE forecasts to the 1-km grid of the LSM used within the WRF-Hydro. Each ensemble member was run through the WRF-Hydro for the full 36-h HRRRE forecast period, followed by an additional 48 h of spindown period comprised of NLDAS-2 forcing with the precipitation field zeroed out.

Streamflow observations were acquired from the USGS database (USGS 2016). The data were available at 15-min intervals and was averaged to 1 h. Time periods with missing data were ignored and not used in the verification to eliminate the possibility of influencing results with interpolated time steps.

b. Site and case selection

The WRF-Hydro was initiated using a 4-yr spinup period and then run from 1 May to 1 November 2018. Model performance during the latter simulation was used to determine which basins to include in this study. This period was chosen due to the occurrence of a wet summer in the study region, leading to several high-flow events for evaluation. Of the original 149 gauge locations contained within the modeled domain, 50 gauges (see Table S1.1 in the online supplemental material) were classified as having acceptable performance based on a Nash–Sutcliffe efficiency greater than 0.4 and percent bias under 40%. We chose a slightly lower threshold for satisfactory model performance than what has been reported in other studies (Moriasi et al. 2007; Madsen et al. 2020) to allow for a larger forecast sample. The selected cases were identified when a flash flood warning, or watch, was issued by the National Weather Service (NWS), or if flooding was observed regardless of a prior advisory issuance (see Table S1.2).

c. Shifting QPF

Kiel et al. (2022) calculated the spatial displacements of predicted precipitation in the HRRRE forecasts by comparing the centroid locations of the precipitation systems in the QPFs to the centroid locations of precipitation systems in the Multisensor Precipitation Estimator (Seo and Breidenbach 2002). Displacements were calculated for the 0–18-h accumulation period and at the hour of CI. The 30 cases used in the Kiel et al. (2022) study were from May to September 2018 and are the same as those used here. It is acknowledged that Kiel et al. (2022) represents a limited period of record, and by using the same forecasts, we are effectively testing our shifting technique with a near-perfect climatology of the forecast system errors. As such, this work represents an ideal scenario.

A skewed-normal random number generator (snRNG) was used to select locations for random shifts of QPF within our model grid that fit the distribution of the HRRRE displacements found in Kiel et al. (2022). Latitude and longitude were sampled separately when running the snRNG because, although they shared similarities in spread, their skew values had differing signs and the paired distributions had almost no correlation.

The method presented in Carlberg et al. (2020) resulted in an 81-member ensemble and thus required 81 separate hydrologic model runs. An ensemble of that size may be impractical in operations; therefore, we first determined the number of shifts (samples) that would be needed to reproduce the distribution of the HRRRE displacement errors with the goal of reducing the number of model forecasts per ensemble to the extent deemed reasonable. Distributions of latitude and longitude displacements were created by sampling the Kiel et al. (2022) 0–18-h accumulation displacements for each HRRRE member using the snRNG. The smallest ensemble tested had three shifts per parent HRRRE member, creating a 36-member ensemble (27 shifted members and 9 nonshifted members), while the largest ensemble tested had eight shifts per member, creating an 81-member ensemble. The snRNG was run 100 times for each sampling regime: 36, 45, 54, 63, 72, and 81 members. For each of the six sampling regimes, the mean absolute deviation (MAD) was calculated for the mean, standard deviation, and skew to evaluate how well the sampled ensemble

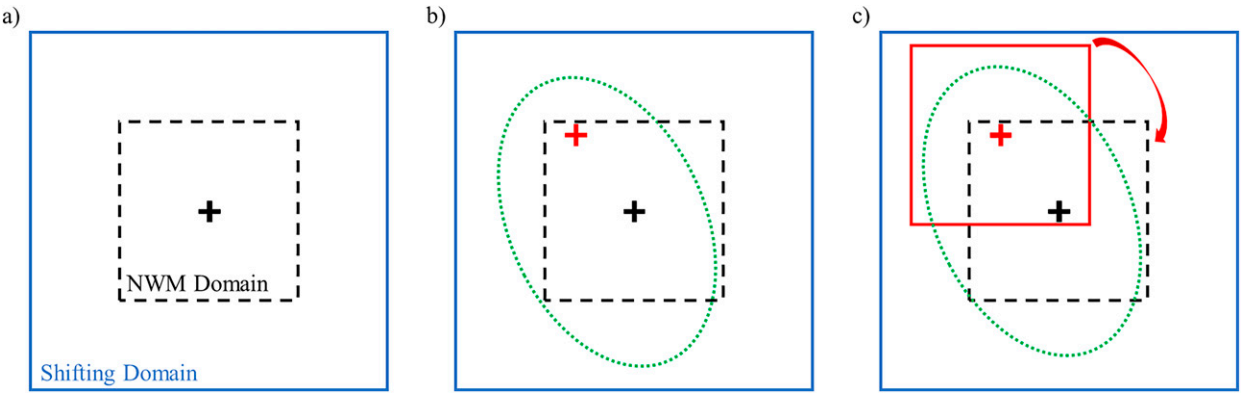

FIG. 2. A visual example of the methodology for shifting QPFs within the WRF-Hydro framework. (a) The solid blue square shows the shifting domain. The black dashed square is the NWM domain. The black cross is the center point of the NWM domain. (b) The green dashed oval represents the climatological displacement distributions found by Kiel et al. (2022). The snRNG picks 54 latitude and longitude points from inside the area of the dashed oval. Those latitude and longitude points then become the centers of a shifted domain (red cross). (c) Data that match the dimensions of the NWM domain, centered at the red cross, are then moved to the center of the NWM domain (black cross).

distribution matched the observed distribution. As the number of shifts in the sampling regime increased, the error between the ensemble distribution and the observed distribution decreased. However, we estimated that the amount of time to preprocess the QPF and run the ensemble for each event would increase by about 25% for every increase of 9 shifted members in the sampling regime. To minimize run time while maintaining a reasonable representation of displacement errors in the ensemble, the 63-member ensemble (Ens63) was chosen. It was comprised of 54 shifted members plus the 9 original nonshifted HRRRE members. The original HRRRE containing only the 9 nonshifted members (Ens9Orig) was used as a point of comparison.

During the regridding process for the HRRRE QPF, data were mapped to the NWM domain, as well as a larger "shifting" domain that fully encompassed the NWM domain but with considerable buffer on every side (Fig. 2). The larger domain was needed to provide a source of QPF for the shifted QPF forcing, while the smaller NWM domain provided a spatial template for data to be shifted into. In the case of a northwest shift (Figs. 2b,c), data matching the grid dimensions of the NWM domain would be taken from the shifting domain and passed to the center of the NWM domain. This acted to shift southeastward the precipitation that was originally to the northwest into the study domain. During rain events, short-term streamflow processes are primarily determined by the precipitation flux. Therefore, only the QPF was shifted and other forecasted input fields (e.g., wind, radiation) that would have a significantly smaller impact on results remained unshifted.

d. Ensemble weighting

The primary weighting scheme for streamflow ensembles uses equal weighting of all members (EqWt). Additional weighting schemes were investigated making use of spatial displacements present at CI. Kiel et al. (2022) hypothesized that displaced rainfall at CI could offer insight into the likely displacement of the 0–18-h accumulation associated with a rainfall system. If this relationship were true, the displacement observed at the beginning of a storm's life cycle could be used to update the ensemble, weighting those shifts that were more likely to result in the smallest spatial error of the total rainfall from the system during its lifetime. To test this hypothesis, two additional ensemble weighting schemes were evaluated. In the first of these schemes, weights were assigned to model members as a function of the inverse of the distance between the centroid of the displaced precipitation at CI and the centroid of the shifted QPF for the rainfall during the lifetime of the system (DistWt). The last weighting scheme leveraged the observation that spatial displacement errors at CI were larger than those for the accumulated 0–18-h QPFs (Kiel et al. 2022). Kiel et al. (2022) showed that the best improvement in the 0–18-h QPF location errors was obtained by adjusting the CI displacement as a function of the directional quadrant in which the CI displacement occurred (i.e., northwest, northeast, southeast, southwest). Thus, the final weighting scheme used the same inverse distance weighting technique as DistWt; however, this approach applied the quadrant-based correction to adjust the centroid of the precipitation system at CI before finding the inverse of the distance (CorrDistWt). Weights for DistWt and CorrDistWt were normalized so the sum of all weights equaled 100%.

The weighting system produced an infinite weight if the distance between centroids was 0 km. In those cases, the zero distance was replaced with an arbitrary value of 0.5 km. The value of 0.5 km avoids the undefined division by 0, while still assigning the most accurate member(s) the highest possible weight. Because the weights go through a process of normalization where each distance between centroids is divided by the sum of the distances among all ensemble members, the results are not sensitive to this change.

The concepts for weighting schemes DistWt and CorrDistWt were also applied to create an additional ensemble for comparison. This ensemble was comprised of a small 9-member

TABLE 1. Ensembles tested in this study.

Abbreviation	Ensemble	Weighting/selection scheme
Ens9Orig	Original HRRRE	Equal weighting of all members
Ens9Orig-DistWt	Original HRRRE	Inverse distance weighting based on observed QPF displacements at CI
Ens9Orig-CorrDistWt	Original HRRRE	Kiel et al. (2022) correction applied to the inverse distance weighting based on observed QPF displacements at CI
Ens63	Original HRRRE with 54 randomly shifted members	Equal weighting of all members
Ens63-DistWt	Original HRRRE with 54 randomly shifted members	Inverse distance weighting based on observed QPF displacements at CI
Ens63-CorrDistWt	Original HRRRE with 54 randomly shifted members	Kiel et al. (2022) correction applied to the inverse distance weighting based on observed QPF displacements at CI
Ens9Sel-DistWt	Members of Ens63 with lowest displacement error grouped by HRRRE member	Members selected based on the distance between (non)shifted member QPF and observed displacement of QPF at CI
Ens9Sel-CorrDistWt	Members of Ens63 with lowest displacement error grouped by HRRRE member	Members selected based on the distance between centroid (non)shifted member QPF and centroid of observed QPF displacement at CI with the Kiel et al. (2022) correction factored in

subset of members selected based on the location of their shifted centroids. To do this, the ensemble members from Ens63 were grouped by their parent HRRRE model member. Then, one member was selected from each parent model group that had the shortest distance between the centroid of its QPF for the rainfall during the lifetime of the system and the centroid of the displaced precipitation at CI (this method is referred to as Ens9Sel). Ens9Sel was meant to keep the variability of each of the perturbation-driven members of Ens9Orig while isolating members that performed best in their depiction of precipitation location at CI. After the selection process is complete, each member within Ens9Sel is given equal weight. A fourth ensemble using a similar method of member selection was tested where the process was repeated without members being grouped by parent model member. After evaluation, it was determined that performance of the fourth ensemble and the Ens9Sel were similar enough that inclusion of both methods in the results was not warranted.

Ens9Sel is only available for comparison when DistWt and CorrDistWt are used because of its dependence on the CI displacement errors for the selection of members. Ens9Orig and Ens63 are compatible with all three weighting approaches. In all, three ensembles using three weighting/selection approaches were tested (Table 1).

e. *Forecast verification*

We use a suite of common dichotomous and probabilistic forecast verification metrics to evaluate the performance of the ensembles for predicting peak discharge and timing. Figure S1.3 shows an example of an ensemble streamflow forecast with flood categories depicted. Ranked histograms (Anderson 1997; Hamill and Colucci 1997; Talagrand et al. 1997) are used to measure how well an ensemble's spread of the forecast represents the true variability of the observations. In this work, the ensemble members are first sorted by the magnitude of peak discharge. The observed peak discharge is compared to the output of the sorted ensemble members, and the position of the observation relative to the ensemble members is determined. That process was repeated for all gauges and cases. The observed frequency of where the observations fell relative to the ensemble peak discharge is plotted using histograms along with a horizontal line representing the uniform distribution of observations. If the histogram forms a "U" shape relative to the horizontal line, the ensemble is considered to be underdispersive. If the histogram forms a bell shape toward the central members, it is considered to be overdispersive. Ideal ensemble performance is shown if/when each column within the histogram matches the horizontal line. The goal of the ranked histogram is to diagnose ensemble consistency. In a consistent ensemble, an observation should be equally likely to fall into any of the ranks within the range provided by the ensemble members (Wilks 2011). Because ranked histograms consider discharge, not probability, they will not be impacted by member weighting. Therefore, the ranked histograms only apply to Ens9Orig, Ens63, Ens9Sel-DistWt, and Ens9Sel-CorrDistWt.

The ranked probability score (RPS) for any individual forecast is the sum of the squared differences of the cumulative distributions of the forecast Y_m and the observed events O_m (Wilks 1995; Franz et al. 2003):

$$\mathrm{RPS} = \sum_{m=i}^{J} (Y_m - O_m)^2. \tag{1}$$

The cumulative distribution of the forecast is

$$Y_m = \sum_{j=1}^{m} y_j,\ m = 1, \ldots, J, \tag{2}$$

where y_j is the relative probability of the forecast, and J is the number of forecast categories (Wilks 1995). The cumulative distribution of the observed streamflow O_m is

$$O_m = \sum_{j=1}^{m} o_j,\ m = 1, \ldots, J, \tag{3}$$

where the category in which the observed discharge peak o_j occurs is given a value of 1, and all categories greater than o_j also receive a value of 1, with categories less than o_j assigned a 0. Smaller scores mean that there is less difference between the forecast probability and the observed probability. The RPS of the magnitude of peak discharge (RPS_Q) was calculated using the following flood categories: less than 50% of action stage, 50% of action stage to action stage, action stage to minor flood stage, minor flood stage to moderate flood stage, moderate flood stage to major flood stage, and greater than major flood stage. [The thresholds defining flood categories for each gauge were obtained from https://water.weather.gov/ahps/ (NOAA 2020) and were converted to units of discharge using the rating curves available at waterwatch.usgs.gov.] For the purposes of this study, we defined the additional threshold of 50% action stage to separate predictions of very low flow from those of moderately high flow. Because major stage was not defined for some gauges, RPS_Q could only be calculated at 37 of the 50 gauges. The RPS_Q was calculated for flood categories, rather than continuous discharge, because the flood categories are relevant in the context of applying forecast information for emergency management and response.

A 2 × 2 contingency table, similar to the setup of Carlberg et al. (2020), was employed to evaluate the forecast skill for prediction of flood or no flood, where "flood" is defined as greater than or equal to minor flood stage. Contingency tables are commonly used to evaluate forecast skill in meteorological forecasting and allow for the evaluation of a simple forecast of a single categorical outcome, given a singular predictand without the consideration of uncertainty (Wilks 1995). Of the 50 gauges that met our standards for Nash–Sutcliffe efficiency and percent bias, only 43 had a defined minor stage. If flooding was observed, and the model predicted flooding, the result is a hit H. Otherwise, if flooding was not predicted by the model and flooding was observed, it is assigned a miss M. Similarly, when flooding was not observed and the model predicted flooding, it is assigned a false alarm (FA). Although each case was selected based on flooding occurring in the region, flood-producing rainfall amounts at each basin did not always occur. If flooding was not observed and the model had not predicted flooding, it is considered a correct negative (CN).

There are several forecast metrics that can be calculated using the information in the 2 × 2 contingency table: probability of detection (POD), the ratio of the number of hits to the total number of events observed,

$$\mathrm{POD} = \frac{H}{H + M}; \tag{4}$$

false-alarm ratio (FAR), the ratio of the number of false alarms to the total number of events forecasted to occur,

$$\mathrm{FAR} = \frac{\mathrm{FA}}{\mathrm{FA} + H}; \tag{5}$$

and equitable threat score (ETS; also called Gilbert skill score), the ratio of hits to hits, misses, and false alarms, with an addition of an estimation for chance to correct for the number of hits that may have occurred due to a chance forecast,

$$\mathrm{ETS} = \frac{H - \mathrm{chance}}{H + \mathrm{FA} + M - \mathrm{chance}}; \tag{6}$$

where chance is calculated as the events forecasted multiplied by the events observed divided by the total number of forecasts N,

$$\mathrm{chance} = \frac{(H + \mathrm{FA}) \times (H + M)}{N}. \tag{7}$$

Probability of exceedance (POE) thresholds of >0%, ≥10%, ≥20%, ≥30%, ≥40%, ≥50%, ≥60%, ≥70%, ≥80%, and ≥90% were used to form 10 contingency tables to assess ensemble forecasting at differing levels of ensemble certainty. POE is the percentage of model members that had forecasted peak streamflow at or above minor flood stage. It should be noted that because Ens9Orig only contains nine members, the lowest possible POE is already greater than 10% (1/9 ~ 11%) for EqWt, as well as Ens9Sel-DistWt and Ens9Sel-CorrDistWt. Although the equally weighted 9-member ensembles would not have a probability less than 10%, 10 levels of probability were used to be consistent with other studies.

In addition to the probabilistic measures mentioned previously, reliability was calculated using forecast probability intervals of 0%–5%, 5%–15%, 15%–25%, 25%–35%, 35%–45%, 45%–55%, 55%–65%, 65%–75%, 75%–85%, 85%–95%, and 95%–100%. Reliability is displayed as a diagram that represents the conditional distribution of observed events given the forecast of an event (Wilks 1995); in this case, the event is the exceedance of minor flood stage. Perfect reliability occurs when the data lie along a 1:1 line; forecasts are overforecasting when they fall to the right of the 1:1 line and under forecasting when they fall to the left of the 1:1 line. To create a summary measure for reliability, MAD provides a way to quantify error as compared to the 1:1 perfect line. MAD for reliability (MAD_{Rel}) is defined here as

$$\mathrm{MAD}_{\mathrm{Rel}} = \frac{\sum_{i=1}^{N} |(F_i - P_i)|}{N}, \tag{8}$$

where F_i is the relative frequency of occurrence of observed events given a forecasted probability within intervals ($i \ldots N$), and P_i is the relative frequency of occurrence of a perfect forecast (thus equal to the forecast probability). Table 2 shows all the abbreviations of the verification metrics as well as their perfect score.

Flood peak timing was analyzed to see if the QPF shifting would lead to any noticeable changes in the temporal aspect of forecast skill. This was done using RPS for the timing of peak discharge (RPS_T), where the categories used to evaluate RPS_T were calculated by splitting forecasted peaks into 2-h bins in a near-continuous calculation. Action stage was the minimum threshold used to define a flood peak, and any member that did not produce output above action stage was

TABLE 2. List of abbreviations of verification metrics and their perfect score.

Metric	Abbreviation	Perfect score
RPS_Q	Ranked probability score for the magnitude of peak discharge	0
POD	Probability of detection	1
FAR	False-alarm ratio	0
ETS	Equitable threat score	1
MAD_{Rel}	Mean absolute deviation of reliability	0
RPS_T	Ranked probability score for the timing of peak discharge	0

regarded as a miss in the timing analysis, and thus, the probability for those members were removed from the total pool used to form the cumulative probability. This made it so an ensemble forecast with a considerable number of misses would produce a worse RPS_T overall. Misses also occurred when ensemble members' peaks occurred at the beginning or end of the period because the full scope of the flood wave could not be captured. In addition to RPS_T, the error between the timing of peak discharge of the ensemble members and observed peak streamflow was calculated. This was done to look at general trends across the different ensembles because weighting approaches could not be factored in.

The goal of this study was to characterize the general performance of the shifting methods; therefore, results were aggregated across all basins and events. While it may be of interest to know how the methods perform for a given watershed, subregion, or storm, such analysis is hindered by a small number of events for any given watershed. Additionally, forecast skill was analyzed based on watershed size, with the median size of basins in our study (~4000 km^2) serving as the threshold between large and small watersheds. Paired two-tailed t tests were used to judge the statistical significance of the results for all metrics. A 90% confidence interval was used as the threshold for assessing significance.

3. Results and discussion

a. *Probabilistic measures*

Ranked histograms show that all the ensembles are underdispersive (Fig. 3), with the observations frequently falling out or at the extreme ends of the ensembles. There is a tendency for the majority of observations to fall below the lowest value from the Ens63 ensemble, while there is a slightly higher tendency for observations to fall above the highest value of the ensemble for the other three ensembles. The problem of underdispersion improves slightly for Ens9Sel-DistWt and Ens9Sel-CorrDistWt as indicated by the higher

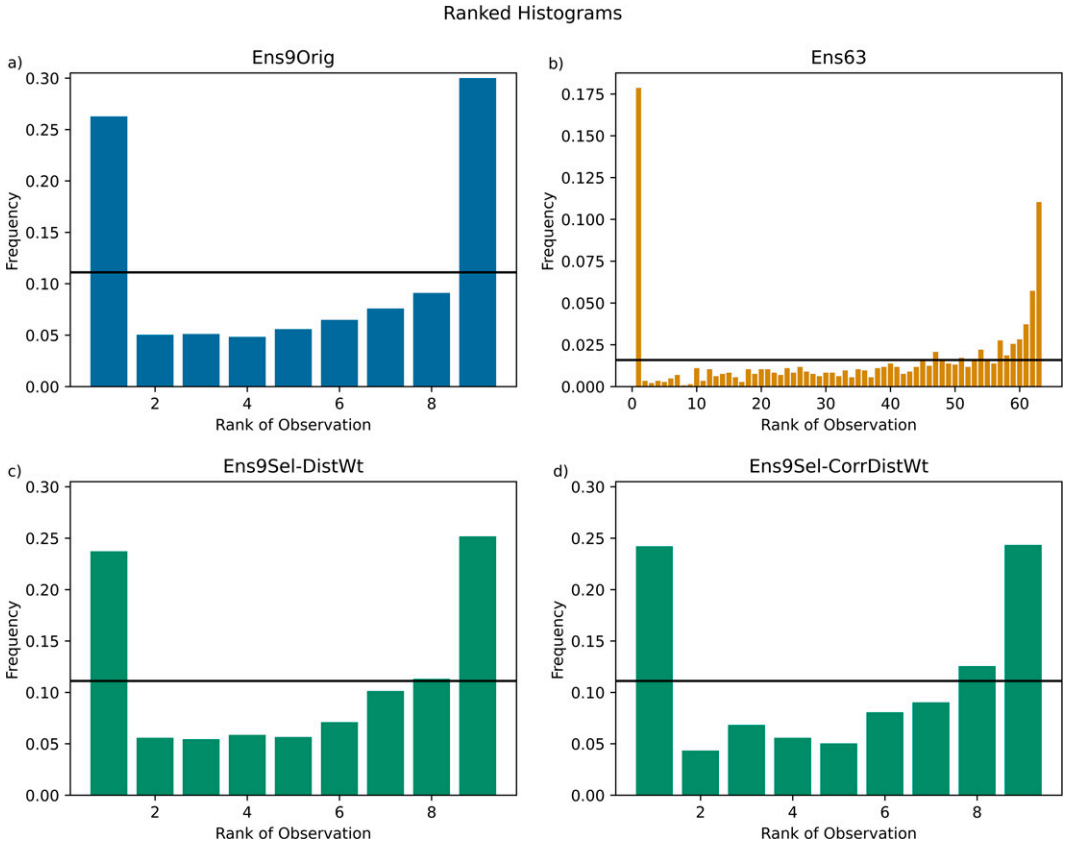

FIG. 3. Ranked histograms for all ensembles (a) Ens9Orig, (b) Ens63, (c) Ens9Sel-DistWt, and (d) Ens9Sel-CorrDistWt. Average case count has been plotted as a horizontal black line to help represent "perfect" reliability.

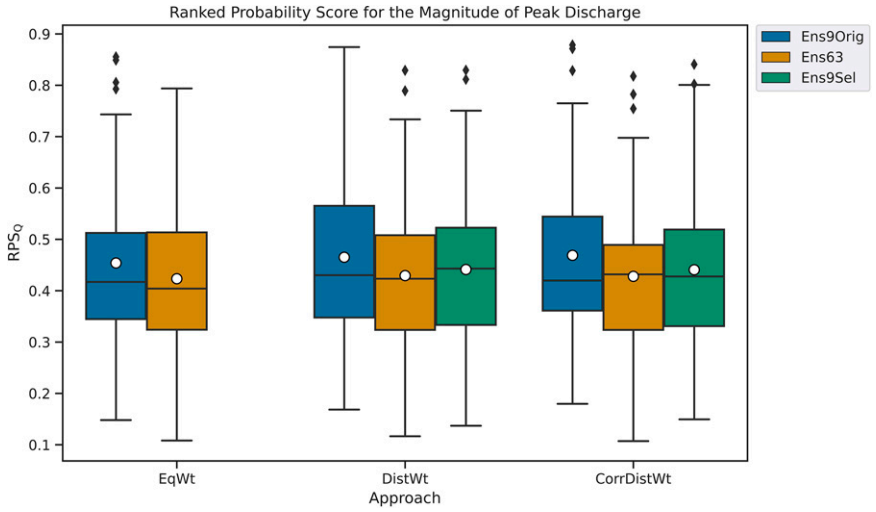

FIG. 4. Box-and-whisker plots for the RPS for RPS$_Q$ for all ensembles (Ens9Orig, Ens63, Ens9Sel) with the three weighting schemes EqWt, DistWt, and CorrDistWt. Means are marked by white circles contained inside the middle two quartiles.

frequency of observations across the middle bins. Moreover, the output of the Ens9Sel-CorrDistWt has slightly better dispersion than the Ens9Sel-DistWt, with the distribution slightly more even across the ranks. In addition to underdispersion, there is a negative bias in all the ensembles, where the observations more heavily favor the higher ranks.

Some of the issues revealed by the ranked histogram can be attributed to the hydrologic model used. The WRF-Hydro tended to overestimate the base flow. In cases where no event occurred, the simulated flow was often already greater than the observation. For these nonevents, the ensembles would be penalized for each member that was overpredicting the base flow (in these cases, there would be little to no rainfall forcing to modify the streamflow response of the model). The problem is exacerbated in the large Ens63 ensemble because it has more members, so the observation more often ends up in the lowest bin of the ranked histogram. Because the accuracy of the baseflow simulation will become a dominant factor in the forecast skill for low-flow, nonflooding situations, bias of simulated discharge was calculated for forecast events where action stage was not met. In those cases, relative bias ranged from 24% to 28%. In contrast, the bias of simulated streamflow for events that surpassed action stage ranged from −5% to 0%.

Hamill (2001) recommends subsetting forecasts to better understand conditional behaviors of ensembles; therefore, ranked histograms were examined for events when flooding did and did not occur. Forecasts for nonflooding events have a positive bias, whereas forecasts for flooding events have a negative bias. Given that model biases were shown to change with flow levels, the conditional biases are likely linked to errors in the hydrologic model in addition to the QPF. The combination of these behaviors for different event types leads to the general pattern of underdispersion seen in Fig. 3.

For RPS$_Q$, Ens63 produced more accurate forecasts on average than Ens9Orig (Fig. 4). The weighting schemes did not improve the RPS$_Q$ scores for either Ens9Orig or Ens63. The 9-member ensemble that used select ensemble members based on their displacement at CI did produce a forecast with lower RPS$_Q$ compared to the Ens9Orig. RPS$_Q$ for Ens9Sel-DistWt and Ens9Sel-CorrDistWt were only slightly worse than Ens63, suggesting that a smaller ensemble could perform as well as the larger ensemble for prediction of flood categories given some information about likely displacements for each member.

Reliability diagrams are shown in Figs. 5a–c for all ensembles and all weighting approaches. The lack of data at 50% reliability for Ens9Orig, as well as Ens9Sel-DistWt and Ens9Sel-CorrDistWt are due to only nine members being in those ensembles. POE values are possible at 44.4% and 55.5% but do not fall within the 45%–55% POE category.

All ensembles overpredict at all levels except the two lowest forecast probability categories, indicating that they assign too much probability compared to the observed frequency. Ens63 has overall better reliability than Ens9Orig. This observation is supported by the lower MAD$_{\text{Rel}}$ (Fig. 5d), with Ens9Orig having a value of 0.223 compared to Ens63 with 0.150. This was the only instance of statistical significance between the MAD$_{\text{Rel}}$ values at a 90% confidence interval.

Similar to RPS$_Q$, the weighting schemes did not significantly improve the verification statistics of the ensembles. Ens9Orig-DistWt and Ens9Orig-CorrDistWt have only slightly better MAD$_{\text{Rel}}$ compared to Ens9Orig, with the improvement occurring in the middle to upper forecast probability range (60%–80%). The weighting schemes reduced the reliability of Ens63. In contrast to the RPS$_Q$, reliability was not improved when the ensemble was reduced to the nine members. Based

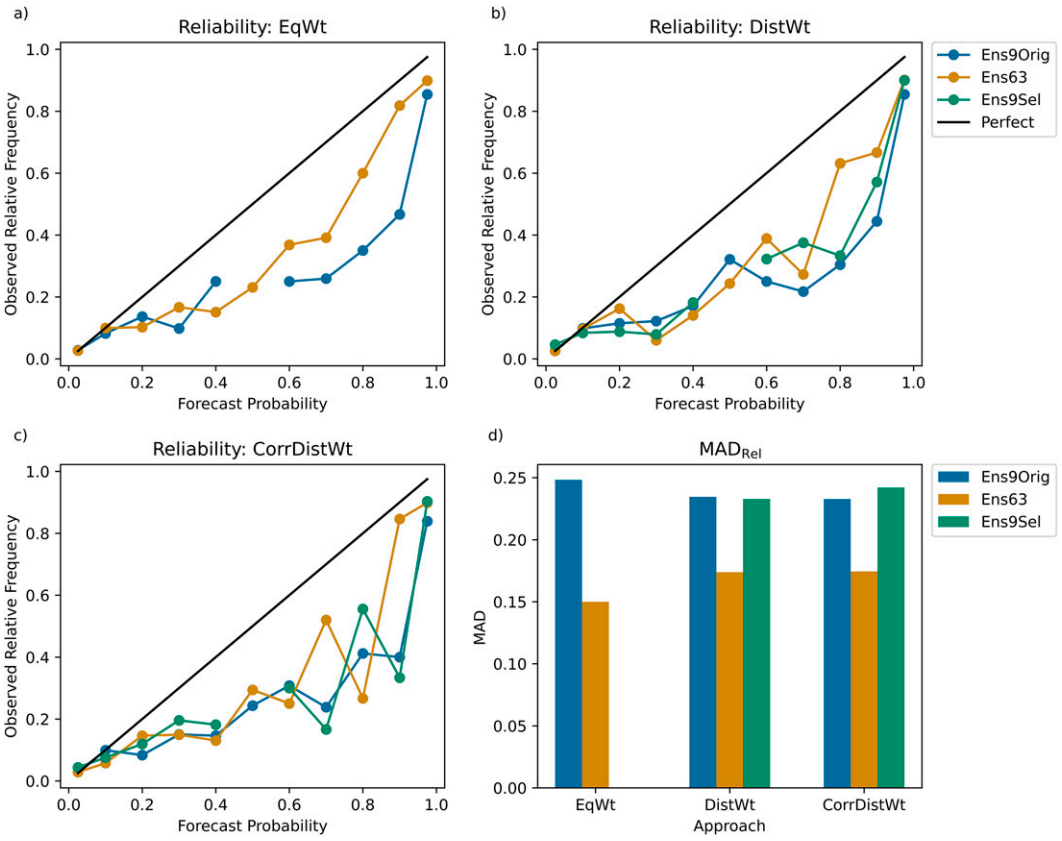

FIG. 5. Reliability curves for forecasts above and below minor flood (colors explained in legend on right) using (a) EqWt, (b) DistWt, and (c) CorrDistWt. A line of theoretically perfect reliability has been added (solid black) to provide a reference. (d) The MAD_{Rel} for all ensembles with all approaches.

on MAD_{Rel}, Ens9Sel-DistWt and Ens9Sel-CorrDistWt did not perform better than the Ens9Orig with these weighting schemes and only slightly better than the equally weighted Ens9Orig.

The lack of perfect reliability is consistent with the issue of underdispersion found in the ranked histograms, which showed that the observations tend to be concentrated at the extreme end of the ensemble range. The overforecasting seen in the reliability diagram also indicates that the members are not adequately distributed across the range of possible outcomes, resulting in probabilities for a given event that are too large on average. The overforecasting in the higher forecast probabilities could stem from the low sample size in these probability ranges.

To summarize the probabilistic verification, the informed spatial shifting of QPF produced improved probabilistic forecasts in terms of both ranked histograms and RPS_Q, where Ens63 outperformed the 9-member ensembles (at a 90% confidence level) for all weighting approaches tested. Reliability matched this pattern, although statistical significance of the better reliability was limited. Moreover, the selection methods used to create Ens9Sel produced better scores than Ens9Orig, while it also appears that there were continued benefits of the increased ensemble size of Ens63.

b. *Dichotomous measures*

POD and FAR for streamflow hitting minor stage or above decreased as POE increased for all ensembles tested (Table 3). Because the general pattern for POD and FAR across POE thresholds was the same for DistWt and CorrDistWt, only results for the equally weighted ensembles are shown. The largest POD was exhibited by Ens63 at the 0% POE threshold. POD drops below 50% at the 80% POE for Ens63 and at 90% POE for Ens9Orig, indicating a slightly better performance for the original HRRRE ensemble. FAR exhibited a similar pattern to POD, with values decreasing as POE thresholds increased. Ens63 had better scores for FAR than Ens9Orig, but the differences were significant only at the 85% confidence level.

Across the POE thresholds, Ens63 had the highest ETS (at 60%) compared to Ens9Orig (Table 4). For POE higher than 60%, Ens9Orig has better ETS, indicating that it produces forecasts of higher skill at higher confidence levels. The weighting schemes improved ETS for the Ens9Orig ensembles, with the highest ETS occurring for Ens9Orig-CorrDistWt at POE of 70%. The weighting schemes improved the Ens63 for POE $>$ 70% but had a negative or mixed impact on ETS for lower POEs. The peak ETS value dropped from 70% POE for the Ens63-DistWt to 50% POE for the Ens63-CorrDistWt. That was the lowest POE threshold associated

TABLE 3. POD and FAR for each POE threshold for Ens9Orig and Ens63 using equal weighting.

	>0%	>10%	>20%	>30%	>40%	>50%	>60%	>70%	>80%	>90%
POD										
Ens9Orig	0.816	0.816	0.76	0.709	0.682	0.62	0.592	0.553	0.514	0.486
Ens63	0.888	0.771	0.659	0.659	0.603	0.575	0.559	0.503	0.475	0.436
FAR										
Ens9Orig	0.65	0.65	0.568	0.517	0.453	0.373	0.312	0.244	0.185	0.13
Ens63	0.762	0.644	0.567	0.494	0.383	0.299	0.248	0.159	0.115	0.093

with a peak ETS outcome of any of the ensembles. This jump between POE thresholds was likely related to the decrease in false alarms. Similarly, the shift in the peak ETS from 60% for Ens9Sel-DistWt to 70% for Ens9Sel-CorrDistWt can be tied to an increase in false alarms and a decrease in hits at the 60% level when the CorrDistWt scheme is applied.

In contrast to the probabilistic measures, the dichotomous measures for peak discharge indicated that Ens9Orig outperformed the other two ensembles and their shifted members. Statistical significance criteria were met for ETS differences between Ens9Orig-CorrDistWt and Ens9Sel-CorrDistWt, where Ens9Orig-CorrDistWt had the better average. The p values for all ensemble comparisons showed no significant differences between any ensembles for DistWt. The highest ETS values occurred for exceedance probabilities greater than 50% (Table 4). Using results from Reed and MacFarlane (2020), who used a 30% exceedance threshold to denote "minor risk" and a 70% threshold for "high risk" for ensemble forecasts, the ensembles tested are most accurate for making statements of higher flood risk. However, there were only a handful of statistically significant differences at the 90% confidence level for results across the various POE levels.

c. *Timing of peak discharge*

The mean timing error between the ensemble forecasts of peak discharge and the observed was 5.15 h early, which is consistent with Towler and McCreight (2021) who found that many of the previous versions of the NWM simulate flood peaks earlier than were observed. There were only minor differences between the ensembles, with means ranging between 4.75 and 5.91 h early. The mode for all four ensembles was identical.

Ens63 had a slightly better mean RPS_T than Ens9Orig (Fig. 6), with the difference being statistically significant (p value of 0.0802). The weighting schemes had minor impacts on the mean RPS_T, and values for DistWt and CorrDistWt were nearly identical. Ens9Sel-DistWt and Ens9Sel-CorrDistWt had the best overall mean RPS_T values. Overall, the differences between the mean RPS_T were within just a few hundredths across the ensembles. The median value and range of results also did not vary greatly. These results show that, although timing of peak discharge can be altered using shifted QPF ensemble members, the impact is generally small.

The outliers seen in Fig. 6 were likely a result of forecast members falling outside the evaluation window with respect to timing of peak, effectively missing the timing of the event by up to 84 h. This most often happened for cases when no precipitation was occurring in the input, and the member had no change in streamflow and no peak discharge. The same phenomena were observed by Carlberg et al. (2020) who found that their methods of shifting QPFs resulted in only a few members producing significant rainfall in the forecast basins, with many members producing little to no change in discharge. The issue with producing many low-flow members remains one of the key challenges to creating ensembles with well-calibrated probability with the shifting methodologies. Postprocessing techniques like those used in Seo et al. (2006)

TABLE 4. ETS for all ensembles (Ens9Orig, Ens63, Ens9Sel) with the three weighting schemes EqWt, DistWt, and CorrDistWt for each POE threshold. The maximum score in each row has been bolded for each ensemble and approach.

	>0%	>10%	>20%	>30%	>40%	>50%	>60%	>70%	>80%	>90%
EqWt										
Ens9Orig	0.221	0.221	0.29	0.321	0.363	0.39	0.409	**0.417**	0.412	0.409
Ens63	0.107	0.22	0.28	0.325	0.375	0.406	**0.419**	0.413	0.404	0.376
DistWt										
Ens9Orig	0.221	0.262	0.301	0.331	0.363	0.385	0.389	0.408	0.406	**0.413**
Ens63	0.107	0.232	0.276	0.323	0.359	0.407	0.409	**0.43**	0.396	0.386
Ens9Sel	0.178	0.178	0.265	0.319	0.367	0.409	**0.415**	0.411	0.414	0.389
CorrDistWt										
Ens9Orig	0.221	0.245	0.292	0.331	0.361	0.387	0.396	**0.437**	0.406	0.402
Ens63	0.107	0.222	0.289	0.337	0.375	**0.421**	0.416	0.416	0.407	0.386
Ens9Sel	0.178	0.178	0.272	0.328	0.359	0.39	0.399	**0.431**	0.403	0.405

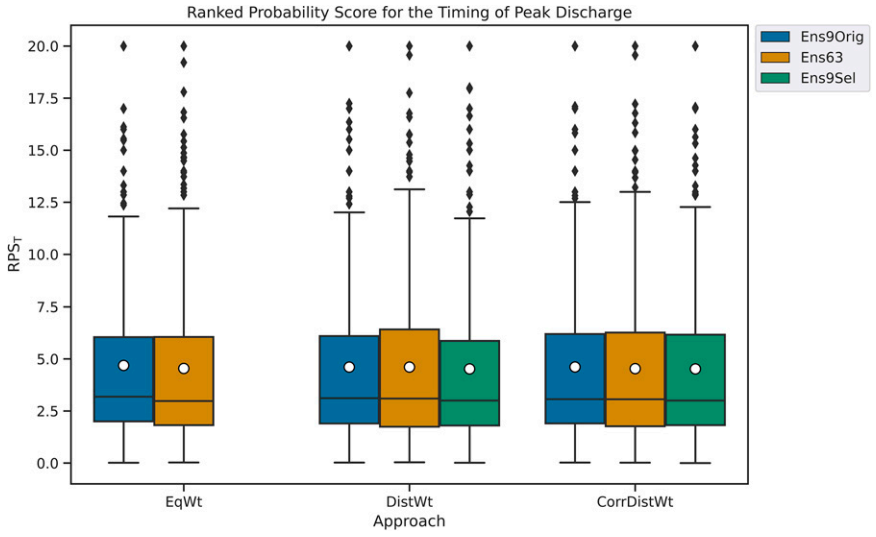

FIG. 6. As in Fig. 4, but with RPS$_T$.

and/or Hashino et al. (2007) could be used to mitigate some of these biases. Likewise, there are other techniques that have been shown to work well in calibration of ensemble predictions in postprocessing (Atger 2003; Gneiting et al. 2005; Brown and Seo 2010).

d. Watershed drainage area analysis

Comparing metrics between the groupings of watersheds, larger watersheds had higher ETS values at all POE thresholds compared to smaller watersheds, with statistically significant results for all three weighting approaches. The same relationship was present for RPS$_Q$ for large basins, with the ensembles producing peak discharge forecasts that more closely matched observations. RPS$_Q$ for large basins was 3.5%–7.5% lower for all of the ensembles compared to the small basins. All differences for RPS$_Q$ were statistically significant at a 90% confidence interval. The evaluation of reliability produced similar results as all four ensembles produced lower MAD$_{\text{Rel}}$ scores for larger watersheds.

The ETS is higher in larger watersheds, with Ens9Orig having only a slightly higher value compared to Ens63 for both large and small watersheds (Table 5). The weighting schemes improved ETS for small basins, but results were more mixed for large watersheds. Results for Ens9Sel were also mixed.

Our results demonstrate that errors in QPF placement are more detrimental to smaller watersheds. Shifted members are more likely to "miss" a small watershed than a large

TABLE 5. Evaluation metric values for small (<4000 km^2) and large watersheds for all ensembles (Ens9Orig, Ens63, Ens9Sel) with the three weighting schemes EqWt, DistWt, and CorrDistWt. The best score for each row and basin size category has been bolded.

	Small watersheds			Large watersheds		
	Ens9Orig	Ens63	Ens9Sel	Ens9Orig	Ens63	Ens9Sel
EqWt						
RPS$_Q$	0.501	**0.49**	—	0.465	**0.434**	—
MAD$_{\text{Rel}}$	0.255	**0.177**	—	0.241	**0.107**	—
ETS	**0.279**	0.236	—	**0.457**	0.454	—
RPS$_T$	3.759	**3.675**	—	4.941	**4.739**	—
DistWt						
RPS$_Q$	0.53	**0.5**	0.513	0.48	**0.437**	0.438
MAD$_{\text{Rel}}$	0.25	**0.189**	0.261	0.224	0.178	**0.177**
ETS	**0.288**	0.242	0.255	0.451	0.46	**0.457**
RPS$_T$	3.707	3.724	**3.645**	4.84	4.782	**4.61**
CorrDistWt						
RPS$_Q$	0.517	**0.501**	0.516	0.479	**0.433**	0.444
MAD$_{\text{Rel}}$	0.241	**0.198**	0.264	0.223	**0.153**	0.195
ETS	**0.29**	0.251	0.259	0.454	**0.458**	0.455
RPS$_T$	**3.729**	3.747	3.828	4.828	4.715	**4.691**

watershed, resulting in more forecast error when events occur, as many members may predict nonevents. The underdispersion seen in the ranked histograms indicates that many of the observations fall at the low end of the ensemble range, and the ensembles are not capturing all possible outcomes. The higher RPS_Q in smaller basins is likely a result of the forecasts having erroneously high probability for nonevents. These findings support the works of Merz et al. (2009), van Esse et al. (2013), Poncelet et al. (2017), and Madsen et al. (2020) in that larger basins tend to score higher in skill metrics and have lower errors.

In contrast to the forecasts of peak discharge, in the case of peak discharge timing, RPS_T was lower for small basins compared to large basins. However, the shifting methods (with and without weighting) did improve the RPS_T values compared to Ens9Orig for large watersheds but not small watersheds. For small basins, Ens63 and Ens9Sel-DistWt were the only instances where ensembles comprised of shifted members produced better RPS_T scores than Ens9Orig. For large basins, Ens9Sel-DistWt had the lowest RPS_T. The shifting methods used in Ens63 and Ens9Sel did not improve forecasts of timing in small watersheds possibly because the streamflow response is more directly a function of the precipitation input. Errors in where the precipitation falls are more likely to propagate to the forecast than in a larger watershed where attenuation processes are more significant and have more opportunity to smooth out or mask errors in location.

When looking at the performance of the ensembles with shifted QPF, there were increases in skill for large basins. The variations in skill between watersheds may also be linked to the morphology and orientation of the basins and the QPF. For example, basins to the west of the Missouri–Mississippi divide tend to have a northeast–southwest orientation, and those to the east tend to have a northwest–southeast orientation. Basins east of the Mississippi River tend to have a northeast–southwest orientation. In general, roughly 50% of the time the QPF (or observed) was aligned southwest–northeast or in scattered cellular precipitation and moved from west to east. When the orientation of the QPF region and the watershed are not aligned, there is less opportunity for the watershed boundary and QPF to overlap, regardless of shifting and/or displacement. For large basins, there is a greater likelihood that some of the QPF may be shifted into the watershed boundaries, whereas, for small basins, again, there is a greater chance of the QPF falling outside the watershed boundaries due to orientation differences.

4. Summary and conclusions

This research examined an ensemble streamflow forecasting method where QPF was shifted in space randomly within a range defined by the climatology of spatial displacement errors in a precipitation ensemble. Our objective was to improve upon prior work by combining the shifting method of Carlberg et al. (2020) with QPF displacement information from Kiel et al. (2022). Our results support findings by Carlberg et al. (2020) that ensembles made up of shifted QPF members provide improved probabilistic streamflow forecasts of flood potential based on RPS_Q, reliability, ETS, and minor improvements in ranked histogram distribution, as compared to ensembles without members driven by shifted QPF. However, for dichotomous forecasts of peak streamflow hitting action stage, there were no statistically significant differences in the performance of the ensembles with and without shifted QPF at several POE thresholds.

Our ensemble member selection method (used to create Ens9Sel) did not consistently improve upon the skill of Ens9Orig and Ens63. Likewise, the member weighting schemes tested also produced mixed results. DistWt showed some increased skill over EqWt depending on the metric used, while CorrDistWt produced marginally lower scores. The correction used in CorrDistWt was meant to reduce the magnitude of displacements present at CI, which are often larger than displacements for the full 0–18-h accumulated precipitation. Although the rain events used in this study had most of their precipitation falling within the first 18 h of the forecast period, and thus should match the 0–18-h climatological spatial displacements found in Kiel et al. (2022) reasonably well, our QPFs included the entire 36-h HRRRE forecast. The Kiel et al. (2022) adjustments may be less relevant beyond forecast hour 18.

In the investigation into the influence of watershed area on ensemble performance, it was found that forecasts for larger watersheds were more skillful than for smaller basins for all metrics except for RPS_T. Because the results were somewhat mixed, the role of the QPF shifting versus the role of the hydrologic model for basins of varying size requires further investigation. When the results were broken down by small and large watersheds, the shifting methods improved the prediction of peak timing for larger watersheds but did not impact overall skill for peak timing.

This work adds additional support for developing an operational framework to address the uncertainty and error introduced into streamflow predictions due to displacement errors in QPF. Work on this topic could be expanded to include other sources of ensemble forecast error such as precipitation timing and/or intensity and hydrologic model error. Further work in characterizing the HRRRE displacement errors beyond forecast hour 18 may provide additional information that could improve the ensemble weighting used in DistWt and CorrDistWt.

These methods have been applied broadly across watersheds of many scales in the north-central United States with all of the cases lumped into a single dataset where available. Some value may be found by examining the effectiveness of this shifting methodology for case studies or individual watersheds while looking at the predictive ability of the ensembles in basins across hydrologic regimes. As mentioned in the results, continued work on this topic could examine the relationships between watershed shape and orientation, QPF shape and orientation, and forecast skill, which may lead to additional avenues for forecast correction and generating guidance in real time. To better understand the potential influences of model error versus QPF error, specifically on short-range flood modeling, sensitivity testing on different models, and/or configurations of the

same model should be undertaken while keeping QPF ensembles consistent.

Acknowledgments. Funding for this work was provided through the NOAA CSTAR Grant NA17NWS4680005.

Data availability statement. The archive of HRRRE model data and WRF-Hydro model output are stored on the Iowa State University box server and are available by request.

REFERENCES

Adams, T. E., III, and R. Dymond, 2019: The effect of QPF on real-time deterministic hydrologic forecast uncertainty. *J. Hydrometeor.*, **20**, 1687–1705, https://doi.org/10.1175/JHM-D-18-0202.1.

Anderson, J. L., 1997: The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Mon. Wea. Rev.*, **125**, 2969–2983, https://doi.org/10.1175/1520-0493(1997)125<2969:TIODCO>2.0.CO;2.

Andresen, J., S. Hilberg, and K. Kunkel, 2012: Historical climate and climate trends in the midwestern USA. National Climate Assessment Midwest Tech. Input Report, 18 pp., https://glisa.umich.edu/media/files/NCA/MTIT_Historical.pdf.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523, https://doi.org/10.1175//1520-0493(2003)131<1509:SAIVOT>2.0.CO;2.

Ball, J. T., I. E. Woodrow, and J. A. Berry, 1987: A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. *Progress in Photosynthesis Research*, J. Biggins, Ed., Springer, 221–224.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The rapid refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Brown, J. D., and G. B. M. Heuvelink, 2006: Assessing uncertainty propagation through physically based models of soil water flow and solute transport. *Encyclopedia of Hydrological Sciences*, M. G. Anderson and J. J. McDonnell, Eds., John Wiley & Sons, 1181–1195, https://doi.org/10.1002/0470848944.hsa081.

——, and D.-J. Seo, 2010: A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. *J. Hydrometeor.*, **11**, 642–665, https://doi.org/10.1175/2009JHM1188.1.

Carlberg, B., K. Franz, and W. Gallus Jr., 2020: A method to account for QPF spatial displacement errors in short-term ensemble streamflow forecasting. *Water*, **12**, 3505, https://doi.org/10.3390/w12123505.

Collier, C. G., 2007: Flash flood forecasting: What are the limits of predictability? *Quart. J. Roy. Meteor. Soc.*, **133**, 3–23, https://doi.org/10.1002/qj.29.

Dowell, D., 2020: HRRR Data-Assimilation System (HRRRDAS) and HRRRE forecasts. NOAA, 8 pp., https://rapidrefresh.noaa.gov/internal/pdfs/2020_Spring_Experiment_HRRRE_Documentation.pdf.

——, C. Alexander, T. Alcott, and T. Ladwig, 2018: HRRR Ensemble (HRRRE) guidance: 2018 HWT spring experiment. NOAA, 6 pp., https://rapidrefresh.noaa.gov/internal/pdfs/2018_Spring_Experiment_HRRRE_Documentation.pdf.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin. *J. Hydrometeor.*, **4**, 1105–1118, https://doi.org/10.1175/1525-7541(2003)004<1105:VONWSE>2.0.CO;2.

Fritsch, J. M., R. J. Kane, and C. R. Chelius, 1986: The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. *J. Climate Appl. Meteor. Climatol.*, **25**, 1333–1345, https://doi.org/10.1175/1520-0450(1986)025<1333:TCOMCW>2.0.CO;2.

Gallus, W. A., Jr., 2010: Application of object-based verification techniques to ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 144–158, https://doi.org/10.1175/2009WAF2222274.1.

Gallus, W. A., Jr., 2012: The challenge of warm-season convective precipitation forecasting. *Rainfall Forecasting*, T. S. W. Wong, Ed., Nova Science Publishers, 129–160.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Gochis, D. J., and Coauthors, 2020: The NCAR WRF-Hydro Modeling System technical description. NCAR Tech. Note, 107 pp., https://ral.ucar.edu/sites/default/files/public/projects/wrf_hydro/technical-description-user-guide/wrf-hydro-v5.1.1-technical-description.pdf.

Goenner, A. R., K. J. Franz, W. A. Gallus Jr., and B. Roberts, 2020: Evaluation of an application of probabilistic quantitative precipitation forecasts for flood forecasting. *Water*, **12**, 2860, https://doi.org/10.3390/w12102860.

Haberlie, A. M., and W. S. Ashley, 2019: A radar-based climatology of mesoscale convective systems in the United States. *J. Climate*, **32**, 1591–1606, https://doi.org/10.1175/JCLI-D-18-0559.1.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and S. J. Colucci, 1997: Verification of ETA–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hapuarachchi, H. A. P., Q. J. Wang, and T. C. Pagano, 2011: A review of advances in flash flood forecasting. *Hydrol. Processes*, **25**, 2771–2784, https://doi.org/10.1002/hyp.8040.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci.*, **11**, 939–950, https://doi.org/10.5194/hess-11-939-2007.

Hejazi, M. I., and M. Markus, 2009: Impacts of urbanization and climate variability on floods in northeastern Illinois. *J. Hydrol. Eng.*, **14**, 606–616, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000020.

Kiel, B. M., W. A. Gallus Jr., K. J. Franz, and N. Erickson, 2022: A preliminary examination of warm season precipitation displacement errors in the upper Midwest in the HRRRE and HREF ensembles. *J. Hydrometeor.*, **23**, 1007–1024, https://doi.org/10.1175/JHM-D-21-0076.1.

Lin, C., S. Vasić, A. Kilambi, B. Turner, and I. Zawadzki, 2005: Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophys. Res. Lett.*, **32**, L14801, https://doi.org/10.1029/2005GL023451.

Madsen, T., K. Franz, and T. Hogue, 2020: Evaluation of a distributed streamflow forecast model at multiple watershed scales. *Water*, **12**, 1279, https://doi.org/10.3390/w12051279.

Merz, R., J. Parajka, and G. Blöschl, 2009: Scale effects in conceptual hydrological modeling. *Water Resour. Res.*, **45**, W09405, https://doi.org/10.1029/2009WR007872.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, 2007: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE*, **50**, 885–900, https://doi.org/10.13031/2013.23153.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

NOAA, 2020: Water. National Weather Service, accessed 26 March 2020, https://water.weather.gov/ahps/.

Poncelet, C., R. Merz, B. Merz, J. Parajka, L. Oudin, V. Andréassian, and C. Perrin, 2017: Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. *Water Resour. Res.*, **53**, 7247–7268, https://doi.org/10.1002/2016WR019991.

Reed, S. M., and A. MacFarlane, 2020: Validation of NWS Hydrologic Ensemble Forecast Service (HEFS) real-time products at the Middle Atlantic River Forecast Center. *34th Conf. on Hydrology*, Boston, MA, Amer. Meteor. Soc., 1A.4, https://ams.confex.com/ams/2020Annual/webprogram/Paper363657.html.

Rezacova, D., Z. Sokol, and P. Pesice, 2007: A radar-based verification of precipitation forecast for local convective storms. *Atmos. Res.*, **83**, 211–224, https://doi.org/10.1016/j.atmosres.2005.08.011.

Seo, B.-C., F. Quintero, and W. F. Krajewski, 2018: High-resolution QPF uncertainty and its implications for flood prediction: A case study for the eastern Iowa flood of 2016. *J. Hydrometeor.*, **19**, 1289–1304, https://doi.org/10.1175/JHM-D-18-0046.1.

Seo, D.-J., and J. P. Breidenbach, 2002: Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *J. Hydrometeor.*, **3**, 93–111, https://doi.org/10.1175/1525-7541(2002)003<0093:RTCOSN>2.0.CO;2.

——, H. D. Herr, and J. C. Schaake, 2006: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci. Discuss.*, **3**, 1987–2035, https://doi.org/10.5194/hessd-3-1987-2006.

Shaw, R. H., and P. J. Waite, 1964: The climate of Iowa: III. Monthly, crop season and annual temperature and precipitation normals for Iowa. Special Rep. 38, 32 pp., https://lib.dr.iastate.edu/specialreports/48/.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) available in the Weather Research and Forecasting (WRF) model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Smith, A. B., 2020: U.S. billion-dollar weather and climate disasters, 1980–present (NCEI Accession 0209268). NOAA/NCEI, accessed 17 March 2020, https://doi.org/10.25921/STKW-7W73.

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. *Wea. Forecasting*, **29**, 894–911, https://doi.org/10.1175/WAF-D-13-00061.1.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, Reading, United Kingdom, ECMWF, 1–25, https://www.ecmwf.int/en/elibrary/12555-evaluation-probabilistic-prediction-systems.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

Towler, E., and J. L. McCreight, 2021: A wavelet-based approach to streamflow event identification and modeled timing error evaluation. *Hydrol. Earth Syst. Sci.*, **25**, 2599–2615, https://doi.org/10.5194/hess-25-2599-2021.

USGS, 2016: USGS Water Data for the Nation. Accessed 24 March 2020, http://waterdata.usgs.gov/nwis/.

van Esse, W. R., C. Perrin, M. J. Booij, D. C. M. Augustijn, F. Fenicia, D. Kavetski, and F. Lobligeois, 2013: The influence of conceptual model structure on model performance: A comparative study for 237 French catchments. *Hydrol. Earth Syst. Sci.*, **17**, 4227–4239, https://doi.org/10.5194/hess-17-4227-2013.

Viterbo, F., and Coauthors, 2020: A multiscale, hydrometeorological forecast evaluation of National Water Model forecasts of the May 2018 Ellicott City, Maryland, flood. *J. Hydrometeor.*, **21**, 475–499, https://doi.org/10.1175/JHM-D-19-0125.1.

Wilks, D. S., 1995: *Statistical Methods in Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

——, 2011: On the reliability of the ranked histogram. *Mon. Wea. Rev.*, **139**, 311–316, https://doi.org/10.1175/2010MWR3446.1.

Xia, Y., and Coauthors, 2012a: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.*, **117**, D03109, https://doi.org/10.1029/2011JD016048.

——, and Coauthors, 2012b: Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *J. Geophys. Res.*, **117**, D03110, https://doi.org/10.1029/2011JD016051.